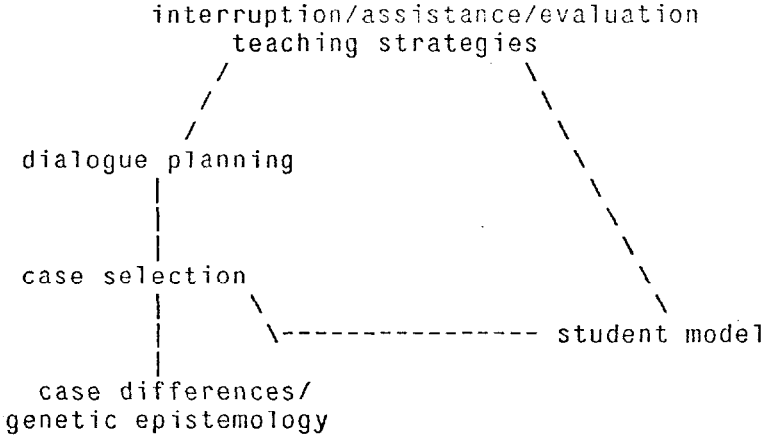
constraints. We are just beginning work, in conjunction with IBM Scientific Labs, to develop an EMYCIN consultation package for electronic fault diagnosis.

GUIDON

A plan for further development of GUIDON is described in terms of a partial ordering of research problems. Improving the student model will receive priority.



Implementation of the strategical methods is now proceeding. There are several tasks (corresponding to the managerial and operational considerations) organized hierarchically. These tasks will be expressed in rule form (if <proc> then <task>).

Structural knowledge will serve to hook these domain independent strategical rules into a particular rule set like MYCIN's. This will involve adding a taxonomic problem classification to the knowledge base and regrouping rules and parameters according to this classification.

Besides using the strategical model for guiding a dialogue with a student, we are investigating the possibility of reconfiguring MYCIN's rule set so that the strategy rules direct a consultation. The result will be a knowledge base of rules and parameters, just like MYCIN's, that does hypothesis formation with focusing by the same backward chaining interpreter we have always used. Even without this step, by formalizing (on paper) a strategical model in terms of production rules, we are led to conclude that it is the exhaustive, depth-first character of MYCIN's search that is different from hypothesis formation, not backward chaining. The strategical rules are meta-rules that modify MYCIN's search. Subgoaling by backward chaining of rules is compatible with both depth-first search and hypothesis formation.

Missing knowledge aside, we find that many of MYCIN's rules are too detailed to be learned by people. We find that people just don't think about the fine-line, statistically-based distinctions that MYCIN rules record. We have developed a way to encode what an expert actually knows by

overlaying qualifications on top of MYCIN's rules. This takes the form of a functional statement (e.g., csf-protein is proportional to intensity and duration of illness) and ranges of discrimination (<100 means viral; >250 means chronic or bacterial; otherwise "it could be anything"). These summary statements capture what the student should learn; they will be used in quizzes based on the rules, as well as for selecting cases.

In a related development, we are trying to record aphorisms and mnemonics that experts use for remembering strategical and mechanistic principles, e.g., "when you hear hoof beats think of horses, not zebras" and "csf glucose is low for bacterial meningitis because bacteria eat the glucose for food" (this is wrong, but physicians remember it and generally don't realize or care that it is wrong!). We find that causal knowledge in our domain serves as a cue for remembering associations; actual diagnosis generally occurs at a level higher than causal mechanism.

ONCOCIN

In the three months remaining in the current year, we expect to have completed the PASCAL interface program that will respond to the special keypad on the Datamedia terminal. We also intend to codify the rules for one more chemotherapy protocol (probably oat cell carcinoma of the lung) in order to verify the generality and flexibility of the representation scheme we have devised. In the coming year, our plans include the following:

(1) To develop the software protocols for achieving communication between the PASCAL interface program and the INTERLISP reasoning program.

(2) To coordinate the printing routines needed to produce hardcopy flowsheets, patient summaries, and encounter sheets.

(3) To install the new terminal and hard copy device in the Oncology Day Care Center for final testing and debugging.

(4) To begin offering the ONCOCIN system for use by oncology faculty and fellows in the chemotherapy clinics (three mornings per week) in which most of the lymphoma patients receive their treatment.

(5) To codify and implement additional protocols contingent upon adequate progress with the steps outline above.

Throughout this work we shall continue to relate the requirements of the system we are developing to the underlying artificial intelligence methodologies. We are convinced that the basic science frontiers of AI are best explored in the context of systems for real world use; thus ONCOCIN serves as a vehicle for developing an improved understanding of the issues that underlie other forms of knowledge engineering.

B. Requirements for Continued SUMEX Use

All the work we are doing (EMYCIN, GUIDON, ONCOCIN, plus continued use of the original MYCIN program) is totally dependent on continued use of the SUMEX resource. The programs all make assumptions regarding the computing environment in which they operate, and the ONCOCIN design in particular depends upon proximity to the 20/20 which will enable us to use a 9600 baud interface. Most of us use SUMEX as the only computer on which we work.

In addition, we have long appreciated the benefits of GUEST and network access to the programs we are developing. SUMEX greatly enhances our ability to obtain feedback from interested physicians and computer scientists around the country. Network access has also permitted high quality formal demonstrations of our work both from around the United States and from sites abroad (e.g., Japan, Sweden, Great Britain).

C. Requirements for Additional Computing Resources

The recent acquisition of the 20/20 by SUMEX has been crucial to the growth of our research work, both to insure high quality demonstrations and to enable us to develop a system such as ONCOCIN for real-world use in a clinical setting. As we continue to develop systems that are potentially useful as stand-alone packages (e.g., an exportable EMYCIN), additional small computers would be particularly valuable resources. It is not yet clear which machines are optimal for the LISP-based applications we are developing, and an opportunity to test our systems on several small-to-medium machines would be invaluable and in keeping with our desire to move some of the AIM products into a community of service users.

As we have mentioned, the response time on the main machine continues to be a major problem during the daytime hours, and is beginning to be limiting on occasion in the evenings as well. Any acquisitions that would provide additional cycles or permit off-loading of some users from the PDP-10 would significantly benefit the SUMEX research community.

The continued growth of our research project, with MYCIN space still required, GUIDON growing, and ONCOCIN now a new and large system, has resulted in some moderate problems with disk allocation as well. We have managed to shuffle allocations reasonably effectively until now, but there is no longer much flexibility and an additional allocation of approximately 2500 pages would greatly relieve the pressure.

D. Recommendations for Future Community and Resource Development

We have two principal recommendations for new SUMEX developments. First, the acquisition of several small machines, linked to the main processor through the ethernet, and each able to run INTERLISP, would allow important experiments in bringing the more mature AIM systems closer to being exportable for use outside of strict research environments.

Second, we propose the formal establishment of a mechanism for providing hardware and communications equipment for SUMEX demonstrations at a distance. There are beginning to be enough invitations for the older AIM systems to be shown at meetings and to funding agencies, that a dedicated system of demonstration equipment and personnel seems appropriate at this time.

9.1.6 Protein Structure Project

Protein Structure Modeling Project

Prof. E. Feigenbaum and Mr. Allan J. Terry
Department of Computer Science
Stanford University

I. Summary of Research Program

A. Technical goals

The goals of the protein structure modeling project are to 1) identify critical tasks in protein structure elucidation which may benefit by the application of AI problem-solving techniques, and 2) design and implement programs to perform those tasks. We have identified two principal areas which are of practical and theoretical interest to both protein crystallographers and computer scientists working in AI. The first is the problem of interpreting a three-dimensional electron density map. The second is the problem of determining a plausible structure in the absence of phase information normally inferred from experimental isomorphous replacement data. Current emphasis is on the implementation of a program for interpreting electron density maps (EDM's).

B. Medical relevance and collaboration

The biomedical relevance of protein crystallography has been well stated in an excellent textbook on the subject (Blundell & Johnson, Protein Crystallography, Academic Press, 1976):

> "Protein Crystallography is the application of the techniques of X-ray diffraction ... to crystals of one of the most important classes of biological molecules, the proteins. ... It is known that the diverse biological functions of these complex molecules are determined by and are dependent upon their three-dimensional structure and upon the ability of these structures to respond to other molecules by changes in shape. At the present time X-ray analysis of protein crystals forms the only method by which detailed structural information (in terms of the spatial coordinates of the atoms) may be obtained. The results of these analyses have provided firm structural evidence which, together with biochemical and chemical studies, immediately suggests proposals concerning the molecular basis of biological activity."

The project involves a collaboration between computer scientists at Stanford University and crystallographers at Oak Ridge National

Laboratories (Dr. Carroll Johnson), the University of California at San Francisco (Dr. Robert Langridge), and the University of California at San Diego (under the direction of Prof. Joseph Kraut). Our principal collaborator at UCSD is Dr. Stephan Freer.

C. Progress summary

We have completed a major cycle of design review and program reorganization, resulting in the system described in publication number three below. The system now has a completely rule-based control structure proceeding from strategy rules, to a set of task rules, ending with individual knowledge sources. This new design seems powerful and flexible enough to provide the basis of a useful EDM interpretation system for protein structure determination.

After building the control structure we wanted, we have worked on building up the knowledge base. Large chunks of knowledge are called "tasks"; we have completed the Initialization task, implemented a tracing task, and implemented a task to split group toeholds. Further details of these tasks and their content can be found in publication number three.

We have also continued our efforts to improve the power of our data representations. Towards this end we have implemented a new preprocessor to assign functional labels to segments. This program consists of heuristics that attempt to capture the knowledge a human uses when he visually examines a skeletonized EDM. We find the use of labeled segments greatly aids the main CRYSALIS program by allowing rules to be written in terms much closer to those which humans use rather than the language in which the EDM skeleton is defined.

Finally, we are compiling documentation on the system and the knowledge it embodies. These documents should be sufficiently complete so that we, or other groups, will have little difficulty picking up where we leave off. We also feel that explicit documentation of our model-building heuristics will be useful to the crystallographic community as it provides a new viewpoint, complementary to traditional crystallographic methods.

The work currently in progress can be characterized as additions to the knowledge base and work on new data representations. Whereas the previously-implemented tracing task attempts to grow an "island of certainty" in the hypothesis in a non-directed manner, we are now working on a task that specifically tries to link two such islands. In addition to this new task, we are augmenting the system's tracing knowledge to deal with small sidechains that seldom appear in the data. The final addition to the knowledge base is an effort to incorporate some notion of stereochemistry and the constraints on three dimensional structure it provides. This will be useful in the matching of features and in the prediction of secondary structure. The last item of work in progress is an attempt to design a data representation that captures volume information. Current representations such as the skeleton preserve topology but do not preserve shape. With the inclusion of volume information, we should be able to capture much of the expert's knowledge of shape and form that presently goes unused.

D. List of Publications

1) Robert S. Engelmore and H. Penny Nii, "A Knowledge-Based System for the Interpretation of Protein X-Ray Crystallographic Data," Heuristic Programming Project Memo HPP-77-2, January, 1977. (Alternate identification: STAN-CS-77-589)

2) E.A. Feigenbaum, R.S. Engelmore, C.K. Johnson, "A Correlation Between Crystallographic Computing and Artificial Intelligence," in Acta Crystallographica, A33:13, (1977). (Alternate identification: HPP-77-15)

3) Robert Engelmore and Allan Terry, "Structure and Function of the CRYSALIS System", Proc. 6IJCAI, 1979. pp250-256 (Alternative identification: HPP-79-16)

4) R. S. Engelmore, A. Terry, S. T. Freer, and C. K. Johnson, "A Knowledge-Based System for Interpreting Protein Electron Density Maps", Abstracts of Amer. Crystallographic Ass. 7,1 (1979) p38

E. Funding status

Grant title: The Automation of Scientific Inference: Heuristic Computing Applied to Protein Crystallography

Principal Investigator: Prof. Edward A. Feigenbaum

Funding Agency: National Science Foundation

Grant identification number: MCS 79-33666

Term of award: December 1, 1979 through November 31, 1981

Amount of award: $35,318 (direct costs only)

II. Interaction with the SUMEX-AIM resource

A. Collaborations

The protein structure modeling project has been a collaborative effort since its inception, involving co-workers at Stanford and UCSD (and, more recently, at Oak Ridge and UCSF). The SUMEX facility has provided a focus for the communication of knowledge, programs and data. Without the special facilities provided by SUMEX the research would be seriously impeded. Computer networking has been especially effective in facilitating the transfer of information. For example, the more traditional computational analyses of the UCSD crystallographic data are made at the CDC 7600 facility at Berkeley. As the processed data, specifically the EDM's and their Fourier transforms, become available, they are transferred to SUMEX via the FTP facility of the ARPA net, with a minimum of fuss. (Unfortunately, other methods of data transfer are often necessary as well

-- see below.) Programs developed at SUMEX, or transferred to SUMEX from other laboratories, are shared directly among the collaborators. Indeed, with some of the programs which have originated at UCSD and elsewhere, our off-campus collaborators frequently find it easier to use the SUMEX versions because of the interactive computing environment and ease of access. Advice, progress reports, new ideas, general information, etc. are communicated via the message and/or bulletin board facilities.

B. Interaction with other SUMEX-AIM projects

Our interactions with other SUMEX-AIM projects have been mostly in the form of personal contacts. We have strong ties to the MYCIN, AGE and MOLGEN projects and keep abreast of research in those areas on a regular basis through informal discussions. The SUMEX-AIM workshops provide an excellent opportunity to survey all the projects in the community. Common research themes, e.g. knowledge-based systems, as well as alternate problem-solving methodologies were particularly valuable to share.

C. Critique of Resource Services

The SUMEX facility provides a wide spectrum of computing services which are genuinely useful to our project -- message handling, file management, Interlisp, Fortran and text editors come immediately to mind. Moreover, the staff, particularly the operators, are to be commended for their willingness to help solve special problems (e.g., reading tapes) or providing extra service (e.g. immediate retrieval of an archived file). We would also like to commend the staff for its extensive help in setting up a link between SUMEX and Dr. Langridge's group at UCSF. Such cooperative behavior is rare in computer centers.

There are several facilities we wish to single out as particularly useful in furthering our research goals. Since the members of the project are physically distant, the MSG program is very useful. Similarly, the file system, the ARCHIVE facility, and the general ease of getting backup files from the operator greatly aid our efforts at coordinating the efforts of collaborators using many large data sets and programs. The crystallographers in the project find SUMEX to be a friendly environment which allows them to do their work with a minimum of dealing with operating system details.

It has become increasingly evident, however, that as CRYSALIS expands, the facility cannot provide enough machine cycles during prime time to support the implementation and debugging of new features. For example, our segment-labeling preprocessor requires about an hour of machine time per 100 residues of protein (this is typically five to eight hours of terminal time during working hours) even when the Lisp code is compiled.

III. Use of SUMEX during the remaining grant period (8/79 - 7/81)

A. Long-range goals

Our short term goals are to build up the knowledge base to the point where it can solve a small, known protein from "live" data. This will probably entail the implementation of about a dozen tasks. By this point we should also have a package of data-reduction programs suitable for export to interested crystallographers.

Our long range goals are the exploitation of the rule-based control structure for investigating alternative problem-solving strategies, the investigation of modes of explanation of the program's reasoning steps, and the expansion and generalization of the system to cover a wider range of input data.

B. Justification for continued use of SUMEX

We feel that SUMEX is the ideal vehicle for further research on CRYSALIS. While some of our work is numerical in nature and uses such facilities as FORTRAN, our main interest is in artificial intelligence. Besides being an expert system of use to the crystallographic community, CRYSALIS is an exploration of the general signal processing problem. We are vitally concerned with issues such as proper architecture for using a wide variety of heuristics effectively and hypothesis formation when both data and model are poor. The utility of our work to the AI community is partially demonstrated by the development of the AGE project, an extension of Ms. Nii's early work on CRYSALIS.

This project progresses by the collaboration of several physically-separated groups. SUMEX provides a unique resource, an electronic community of researchers in our field, through the many systems such as net mail, country-wide access, and community workshops. We feel that CRYSALIS would not be possible outside of such a community.

C. Needs and plans for other computing resources

Our major need for other computing resources is for graphical display of our data and results. This need will be met by use of Dr. Langridge's Evans and Sutherland Picture System at UCSF and Dr. Johnson's raster-based graphics system at ORNL. The major impediment is SUMEX's current inability to support data transfer to other machines at more than 1200 baud. We are attempting to link SUMEX to UCSF by using FTP over the ARPAnet to the LBL machine and then use an existing link from LBL to UCSF.

D. Recommendations for future community and resource development

There are two recommendations we wish to make, the first and most important is to expand the computing power available to SUMEX users. CRYSALIS is an inherently-large problem. Proteins contain hundreds, to thousands of atoms which means large hypothesis structures, large quantities of data, and a compute-bound inference program. As the system grows to maturity, we expect increasingly serious problems with address space limitations and with machine cycle availability.

The second recommendation is that SUMEX develop some relatively inexpensive file transfer facility for machines not on the ARPAnet. Software for this already exists in the form of the TTYFTP program (or possible future programs like it, but in a more portable language), the development needed is in hardware and in the TENEX operating system so that transfer rates greater than 1200 baud can be achieved. We are motivated to recommend this not only by our own need for such a facility, but also by the belief that it would aid other collaborations involving SUMEX and outside computers (the SECS project for example), and aid in the dissemination of useful programs from the research setting of SUMEX to user laboratories.

9.1.7 RX Project

The RX Project: Deriving Medical Knowledge from Time-Oriented Clinical Databases

Robert L. Blum, M.D.
Division of Clinical Pharmacology
Department of Internal Medicine
Stanford School of Medicine

Gio C. M. Wiederhold, Ph.D.
Departments of Computer Science and Electrical Engineering
Stanford University

I. Summary of Research Program

I.A. Technical goals:

Introduction:

Medical and Computer Science Goals

The objective of the RX Project is to develop a medical information system capable of accurately deriving knowledge of the course and consequences of treatment of chronic diseases from a large collection of stored patient records.

Computerized clinical databases and automated medical records systems have been under development throughout the world for at least a decade. Among the earliest of these endeavors was the ARAMIS Project, (American Rheumatism Association Medical Information System) under development at Stanford by Dr. James Fries and his colleagues since 1967. A prototype ambulatory records system was generalized in the early 1970's by Prof. Gio Wiederhold and Stephen Weyl in the form of a Time-Oriented Database (TOD) System. The TOD System, run on the IBM 370/3033 at the Stanford Center for Information Processing (SCIP), now supports the ARAMIS Project as well as a host of other chronic disease databases which store patient data gathered at many institutions nation-wide. At the present time ARAMIS contains records of over 10,000 patients with a variety of rheumatologic diagnoses. Over 30,000 patient visits have been recorded, accounting for 20,000 patient-years of observation.

The fundamental objective of ARAMIS, the other TOD research groups, and all other clinical data bank researchers is to use the raw data which has been gathered by clinical observation in order to study the evolution and medical management of chronic diseases. Unfortunately, the process of reliably deriving knowledge from raw data has proven to be refractory to existing techniques because of problems stemming from the complexity of disease, therapy, and outcome definitions; the complexity of time relationships; complex causal relationships creating strong sources of bias; and problems of missing and outlying data.

A major objective of the RX Project is to explore the utility of symbolic computational methods and knowledge-based techniques at solving this problem of accurate knowledge inference from non-randomized, non-protocol patient records. A central component of RX is a knowledge base of medicine and statistics, organized as a hierarchy or taxonomic tree consisting of nodes with attached data and procedures. Nodes representing diseases and therapeutic regimens contain procedures which use a variety of time-dependent predicates to label patient records in the database, facilitating the retrieval of time-intervals of interest in the records. The database is then inverted so that each node or object in the knowledge base contains pointers to all time-intervals during which its definition is satisfied.

Nodes in the knowledge base also contain lists of other nodes which are causally related. These functional dependencies are used to infer causal pathways among nodes for purposes of selecting confounding variables which need to be controlled for in the study of a specific hypothesis. Causal pathways may also be used in an exploratory mode to discover new hypotheses.

To study a particular causal hypothesis the knowledge base also contains information on the applicability of various statistical procedures and procedures for applying them.

I.B. Medical Relevance and Collaboration

As a test bed for system development our focus of attention has been on the records of patients with systemic lupus erythematosus (SLE) contained in the Stanford portion of the ARAMIS Data Bank. SLE is a chronic rheumatologic disease with a broad spectrum of manifestations which can lead to death in the third decade of life. With many perplexing diagnostic and therapeutic dilemmas, it is a disease of considerable medical interest.

In the future we anticipate possible collaborations with other project users of the TOD System such as the National Stroke Data Bank, the Northern California Oncology Group, and the Stanford Divisions of Oncology and of Radiation Therapy.

The RX Project is a new research effort only in existence for about a year, and, hence the project is very much in a developmental stage. The primary issues being addressed at this stage are those concerned with the specifics of knowledge representation and flow of control, rather than with the testing of specific hypotheses in chronic disease management.

We believe that this research project is broadly applicable to the entire gamut of chronic diseases which constitute the bulk of morbidity and mortality in the United States. Consider five major diagnostic categories which are responsible for approximately two thirds of the two million deaths per year in the United States: myocardial infarction, stroke, cancer, hypertension, and diabetes. Therapy for each of these diagnoses is fraught with controversy concerning the balance of benefits versus costs.

1) Myocardial Infarction: Indications for and efficacy of coronary artery bypass graft vs. medical management alone. Indications for long-term antiarrhythmics ... long-term anticoagulants. Benefits of cholesterol-lowering diets, exercise, etc.

2) Stroke: Efficacy of long-term anti-platelet agents, long-term anticoagulation. Indications for revascularization.

3) Cancer: Relative efficacy of radiation therapy, chemotherapy, surgical excision - singly or in combination. Optimal frequency of screening procedures. Prophylactic therapy.

4) Hypertension: Indications for therapy. Efficacy versus adverse effects of chronic antihypertensive drugs. Role of various diagnostic tests such as renal arteriography in work-up.

5) Diabetes: Influence of insulin administration on microvascular complications. Role of oral hypoglycemics.

Despite the expenditure of billions of dollars over recent years for randomized controlled trials (RCT's) designed to answer these and other questions, answers have been slow in coming. RCT's are expensive of funds and personnel. The therapeutic questions in clinical medicine are too numerous for each to be addressed by its own series of RCT's.

On the other hand, the data regularly gathered in patient records in the course of the normal performance of health care delivery is a rich and largely underutilized resource. The ease of accessibility and manipulation of these data afforded by computerized clinical data banks holds out the possibility of a major new resource for acquiring knowledge on the evolution and therapy of chronic diseases.

The goal of the research which we are pursuing on SUMEX is to increase the reliability of knowledge derived from clinical data banks with the hope of providing a new tool for augmenting knowledge of diseases and therapies as a supplement to knowledge derived from formal prospective clinical trials. Furthermore, the incorporation of knowledge from both clinical data banks and other sources into a uniform knowledge base should increase the ease of access by individual clinicians to this knowledge and thereby facilitate both the practice of medicine as well as the investigation of human disease processes.

Highlights of Research Progress

1 July 1979 to 1 April 1980

Our predominant objective was to detail the overall conceptual framework for the knowledge base and to develop the extensive computational machinery necessary for retrieving, analyzing, and displaying defined time-intervals within patient records.

The RX Knowledge Base (KB):

The central component of RX is a knowledge base of medicine and statistics, organized as a frame-based, taxonomic tree consisting of units with attached data and procedures. Units representing diseases and therapies contain procedures which use a variety of time-dependent predicates to label the patient records, facilitating the retrieval of time-intervals of interest in the records. Other units representing statistical techniques are used to map hypotheses onto study designs and event definitions. Implementing the algorithms and data structures of this KB was one of the major tasks of the current year.

At the current time the RX KB contains about 200 units of which 75 contain definitions and other relevant information pertaining to disease courses, effects of drugs, lab values, etc. This information compromises a small subset of medical knowledge dealing with some of the signs and symptoms of systemic lupus erythematosus (SLE) as well as the effects and indications of some drugs used for this disease. Other units contain machine-readable knowledge of statistical techniques needed for testing entered hypotheses. There are approximately 40 time-dependent functions used to map from the database values onto defined units.

The entire RX system currently contains approximately 250 INTERLISP functions accounting for 75 disk pages of code. The KB is about 30 disk pages. One disk page = 512 words * 36 bits per word. Also one disk page = approx. 1.5 typed pages on 8.5 by 11.5 inch paper.

Statistical Interfaces:

Once the relevant episodes have been defined and retrieved from the database they must be analyzed statistically. In order to do this we use the SPSS package (Statistical Package for the Social Sciences) available on SUMEX. A collection of RX programs create SPSS "source decks" containing card images of the appropriate commands along with the extracted data. RX then calls the operating system and runs SPSS on the source file. The human-readable listing is then searched for important results which are automatically extracted and interpreted.

Time-Oriented Graphics Package:

This package enables data on an individual patient to be graphed over time, either linearly by visit or by calendar time with a "telescoping" capability. The program overlays graphs of both point data and data represented as episodes.

Study Editor:

Dr. Jerrold Kaplan, a research associate affiliated with the project, has implemented an additional package of programs which display to the clinician user those decisions which have been made by the knowledge base concerning which statistical techniques are to be employed, which variables are to be controlled for, and which time intervals are to be excluded. This affords the user with a means for seeing a sketch of the study plan before it is executed, and enables him to modify that plan.

Clinical Study: The Effect of Prednisone on Cholesterol

As a testbed for the prototype system we have been investigating the hypothesis that the steroid, prednisone, produces a significant elevation of plasma cholesterol. To test this hypothesis, the records of 50 patients with systemic lupus erythematosus (SLE) were transferred from the ARAMIS Database to SUMEX. Of these patients, 18 were found to have five or more cholesterol determinations and to have had sufficient variance in their prednisone regimens to be testable. The KB is used to elaborate a complex causal model for the prednisone/cholesterol hypothesis which is tested using a hierarchical multiple regression method with time-lagged values. The KB is used to determine sources of possible bias and to control for those variables in the regression or to eliminate corresponding time-intervals from records. An empirical Bayes method is used to average the estimated effects in patients with varying amounts of data.

The result, a highly statistically significant elevation of cholesterol by prednisone, will be submitted for publication during the coming year.

Research In Progress

Much work remains to be done in expanding the system software and in expanding the knowledge base. Current work is addressed to increasing the flexibility of the time-segmentation functions and enriching the data structures which encode relationships among objects.

We are trying to make increasingly general the class of medical hypotheses which the system can analyze automatically. This requires incorporating knowledge of additional statistical methods into the KB and the development of expanded capabilities for interfacing RX to on-line statistical packages. We are also attempting to generalize our algorithms for selecting the set variables which may potentially confound a given hypothesis. As a means for testing and expanding the system's capabilities we intend to perform several specific studies of importance in the management of the rheumatic diseases. Our study of the effect of prednisone on cholesterol was mentioned above. Other studies now being planned include the effect of chronic aspirin ingestion on liver function in rheumatoid arthritis, the specific incidence of infectious complications of steroids as a function of dose and duration, and the utility of various autoantibodies in the prediction of flares of SLE as compared to the utility of other indicators.

Finally, we are developing a methodology for discovering hypotheses of interest in the database using a heuristically guided search of large matrices of simple and partial correlation coefficients.

Publications

Blum, Robert L.; Wiederhold, Gio: Inferring Knowledge from Clinical Data Banks Utilizing Techniques from Artificial Intelligence. Proc. of The 2nd Annual Symp. on Computer Applications in Medical Care, pp. 303 to 307, IEEE, Washington, D.C., November 5-9, 1978

Blum, Robert L.: Automating the Study of Clinical Hypotheses on a Time-Oriented Data Base: The RX Project. Submitted for publication to MEDINFO80, Tokyo, Japan, Oct. 1980

Wiederhold, Gio: Databases in Healthcare. To be published in a compendium series on Technology in Healthcare, sponsored by the Healthcare Technology Center, Univ. of Missouri, Columbia, Mo., also available as Stanford CS Report 80-790

Funding Support Status

1) A Computer-Based System for Advising Physicians on Clinical Therapeutics
Robert L. Blum, M.D.: Awardee
Post-Doctoral Research Fellowship in Clinical Pharmacology
Pharmaceutical Manufacturers' Association Foundation
Total award: $32,500 (direct)
Term: July 1, 1978 to June 30, 1980

2) Integrating Medical Knowledge and Clinical Data Banks
Robert L. Blum, M.D.: Principal Investigator
National Library of Medicine, New Investigator Award
Total award: $90,000 (direct)
Term: July 1, 1979 to June 30, 1982

3) Integrating Medical Knowledge and Clinical Data Banks
Gio C. M. Wiederhold, Ph.D.: Principal Investigator
National Center for Health Services Research, Small Grants
Total award: $35,000 (direct)
Term: April 1, 1979 to March 31, 1981

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

II.A. Collaborations

Since our project is new, we do not yet have public versions of the programs. There is, however, a large sphere of collaboration which we expect in the future. Once the RX program is developed, we would anticipate collaboration with all of the ARAMIS project sites in the further development of a knowledge base pertaining to the chronic arthritides. The ARAMIS Project at SCIP is used by a number of institutions around the country via commercial leased lines to store and process their data. These institutions include the University of California School of Medicine, San Francisco and Los Angeles; The Phoenix Arthritis Center, Phoenix; The University of Cincinnati School of Medicine; The University of Pittsburgh School of Medicine; Kansas University; and The University of Saskatchewan. All of the rheumatologists at these sites have closely collaborated with the development of ARAMIS, and their interest in and use of the RX project is anticipated. We hasten to mention that we do not expect SUMEX to support the active use of RX as an on-going service to this extensive network of arthritis centers, but we would like to be able to allow the national centers to participate in the development of the arthritis knowledge base and to test that knowledge base on their own clinical data banks.

B. Interactions with Other SUMEX-AIM Projects

Several of the concepts incorporated into the design of the RX Project have been inspired by other SUMEX-AIM Projects. The RX knowledge base is similar to the Units Package of the MOLGEN PROJECT. The production rule inference mechanism used by us is similar to that in the MYCIN Project.

Several programs developed by the MYCIN group are regularly used by RX. These include disk hash file facilities, text editing facilities, and miscellaneous LISP functions. Regular communication on programming details is facilitated by the on-line mail system.

C. Critique of Resource Management:

The SUMEX KI-10 has been severely overloaded for at least a year. Working in LISP is impossible during the day and is even difficult at times which were formerly low utilization times. This has forced us to rely increasingly on other local computation facilities.

The SUMEX resource management, per se, has always been accessible and cooperative in trying to provide our project with adequate resources subject to prevailing constraints.

III. RESEARCH PLANS

The overall goal of the RX Project is to develop a computerized medical information system capable of accurately extracting medical knowledge pertaining to the therapy and evolution of chronic diseases from a database consisting of a collection of stored patient records.

Goals for the year August, 1980 to July, 1981 have been detailed in section IC. above on research in progress. To summarize that section, our main short-term goal is to generalize and refine our methods for labeling and retrieving time-intervals or episodes from individual patient records and to generalize the class of hypotheses which the system is capable of analyzing. This requires further refinements in RX's algorithms for choosing and controlling for variables which may potentially confound an hypothesis of interest.

Long-Range Goals: August, 1981 to July, 1986

There are two inter-related long-range goals of the RX Project: 1) automatic discovery of knowledge in a large time-oriented database and 2) provision of assistance to a clinician who is interested in testing a specific hypothesis. These tasks overlap to the extent that some of the algorithms used for discovery are also used in the process of testing an hypothesis.

We hope to make these algorithms sufficiently robust that they will work over a broad range of hypotheses and over a broad spectrum of data distributions in the patient records.

Justification for Continued Use of SUMEX

Computerized clinical data banks possess great potential as tools for assessing the efficacy of new diagnostic and therapeutic modalities, for monitoring the quality of health care delivery, and for support of basic medical research. Because of this potential, many clinical data banks have recently been developed throughout the United States. However, once the initial problems of data acquisition, storage, and retrieval have been dealt with, there remains a set of complex problems inherent in the task of accurately inferring medical knowledge from a collection of observations in patient records. These problems concern the complexity of disease and outcome definitions, the complexity of time relationships, potential biases in compared subsets, and missing and outlying data. The major problem of medical data banking is in the reliable inference of medical knowledge from primary observational data.

We see in the RX Project a method of solution to this problem through the utilization of knowledge engineering techniques from artificial intelligence. The RX Project, in providing this solution, will provide an important conceptual and technologic link to a large community of medical research groups involved in the treatment and study of the chronic arthritides throughout the United States and Canada, who are presently using the ARAMIS Data Bank through the SCIP facility via TELENET.

Beyond the arthritis centers which we have mentioned in this report, the TOD (Time-Oriented Data Base) User Group involves a broad range of university and community medical institutions involved in the treatment of cancer, stroke, cardiovascular disease, nephrologic disease, and others. Through the RX Project, the opportunity will be provided to foster national collaborations with these research groups and to provide a major arena in which to demonstrate the utility of artificial intelligence to clinical medicine.

SUMEX as a Resource

To discuss SUMEX as a resource for program development, one need only compare it to the environment provided by our other resource, the IBM 370/168 installation at SCIP - the major computing resource at Stanford. Of the programs which we use daily on SUMEX -INTERLISP, MSG, TVEDIT, BBD, LINK- there is nothing even approaching equivalence on the 370, despite its huge user community. These programs greatly facilitate communication with other researchers in the SUMEX community, documentation of our programs, and the rapid interactive development of the programs themselves. The development of a program involving extensive symbolic processing and as large and complex as RX at the SCIP facility, would require a staff many times as large as ours. The SUMEX environment greatly increases the productive potential of a research group such as ours to the point where a large project like RX becomes feasible.

Computation resources required by RX:

Disk Allocation:

RX requires the use of two large data files which need to be kept on-line: the patient database (DB) and the knowledge base (KB). In the course of testing a hypothesis several other files are used: inverted files, source files for statistical processing, LISP SYSOUT files, etc. Our current total disk allocation of 1500 pages for all RX group members has been just adequate. In the future, with anticipated expansions in numbers of patients and size of the KB, we intend to request an increase of our total allocation to 2000 pages.

Programs:

RX is written in INTER-LISP. To increase our useable address space, we actually use a stripped-down version prepared by William VanMelle of the MYCIN Project. To run statistical data RX calls SPSS in an inferior fork. The text editor, TVEDIT, is also called from an inferior exec fork.

Other Computational Resources

It is clear that the scope of potential application of the RX Project is large. Within the term of the SUMEX-AIM grant projected through July, 1986, we anticipate the involvement of several of the national ARAMIS collaborating institutions in developing and testing arthritis knowledge bases which reflect their own patient populations and therapeutic biases. The current SUMEX machine configuration will not be able to support this national interaction because the central processors of the KI-10 are already taxed to the limit. Ours is among the SUMEX groups which would greatly benefit by the addition of one or more PDP-10 compatible machines, which could provide support to our anticipated national user community. Another resource which would be highly desirable is a faster and more reliable means for transferring data interactively between SUMEX and the SCIP IBM 370. Our current method utilizes a 2400 baud line with transmission from SCIP to SUMEX only, and is fraught with a high error rate. The addition of a reliable local network facility would greatly facilitate our ability to transfer patient files from SCIP to SUMEX and to transfer statistical source matrices back to SCIP to be run on that machine.

D. Recommendations for Resource Development:

SUMEX is heavily loaded everyday and almost every evening. Program research is next to impossible during those periods. Program development would be greatly facilitated by the addition of any resources which lessened this loading: upgrading the current machine to a KL or adding core to decrease page swapping.

9.2 National AIM Projects

The following group of projects is formally approved for access to the AIM aliquot of the SUMEX-AIM resource or the Rutgers-AIM resource. Their access is based on review by the AIM Advisory Group and approval by the AIM Executive Committee.

9.2.1 Acquisition of Cognitive Procedures (ACT)

Acquisition of Cognitive Procedures (ACT)

Dr. John Anderson
Carnegie-Mellon University

I. Summary of Research Program

A. Project Rationale:

To develop a production system that will serve as an interpreter of the active portion of an associative network. To model a range of cognitive tasks including memory tasks, inferential reasoning, language processing, and problem solving. To develop an induction system capable of acquiring cognitive procedures with a special emphasis on language acquisition and problem-solving skills.

B. Medical relevance and collaboration:

1. The ACT model is a general model of cognition. It provides a useful model of the development of and performance of the sorts of decision making that occur in medicine.

2. The ACT model also represents basic work in AI. It is in part an attempt to develop a self-organizing intelligent system. As such it is relevant to the goal of development of intelligent artificial aids in medicine.

We have been evolving a collaborative relationship with James Greeno and Allan Lesgold at the University of Pittsburgh. They are applying ACT to modeling the acquisition of reading and problem solving skills. We have made ACT a guest system within SUMEX. ACT is currently at the state where it can be shipped to other INTERLISP facilities. We have received a number of inquiries about the ACT system. ACT is a system in a continual state of development but we periodically freeze versions of ACT which we maintain and make available to the national AI community.

C. Highlights of Research Progress:

This last year has seen developments in two main directions. We are completing developing and documenting a system (ACTF) that is capable of a relatively rich variety of cognitive learning and we are completing an application to the modelling of the acquisition of proof skills in high school students.

Our ACTF system is a production system that operates in a semantic network data base. Our learning work has been focused on ways of increasing the power of production systems for performing various tasks. One class of learning mechanisms concern what we call knowledge compilation. This involves automatic mechanisms for creating productions

that directly perform behavior that formerly required interpretative processing of knowledge in the semantic network. These compilation mechanisms also model the process by which human experts develop special purpose procedures to deal with the different types of problems that occur in their domain of expertise.

Another class of learning mechanisms are concerned with tuning existing procedures so that they apply more appropriately. There are various mechanisms concerned with extending or generalizing the range of application of a procedure. In the past year we have been working at reducing these different generalization processes to a common partial matching process. In addition to generalization, tuning occurs in the ACTF system by means of discrimination and composition. Discrimination is a process for restricting the range of applicability of a production. Composition attempts to build macro-operators out of a series of productions.

The third direction of our learning work has been concerned with developing a flexible strength-based set of conflict resolution rules. Here we are concerned with modelling the gradual improvement seen in human cognitive skills and also providing the system with the resilience so that it can recover from noise and changes in environmental contingencies.

A manual has been under construction describing these changes. We plan to have a final version of the ACTF system by the end of May and the manual should be finished by the end of the summer.

We have been applying this theory in detail to a simulation of how students acquire proof skills in geometry. We have a more or less thorough analysis of how students learn new postulates of geometry; initially use these postulates in an interpretative fashion, integrating them with prior knowledge; how they compile special purpose procedures that directly apply this knowledge to proof generation; and how these procedures become tuned with practice. This application has provided strong evidence for most of the learning developments in the ACT system. It has also forced us to develop formalisms for how planning and problem-solving should be structured within a production-system framework.

D. List of project publications:

[1] Anderson, J.R. Language, Memory, and Thought. Hillsdale, N.J.: L. Erlbaum, Assoc., 1976.

[2] Kline, P.J. & Anderson, J.R. The ACTE User's Manual, 1976.

[3] Anderson, J.R., Kline, P. & Lewis, C. Language processing by production systems. In P. Carpenter and M. Just (Eds.). Cognitive Processes in Comprehension. L. Erlbaum Assoc., 1977.

[4] Anderson, J.R. Induction of augmented transition networks. Cognitive Science, 1977, 125-157.

[5] Anderson, J.R. & Kline, P. Design of a production system. Paper presented at the Workshop on Pattern-Directed Inference Systems, Hawaii, May 23-27, 1977.

[6] Anderson, J.R. Computer simulation of a language acquisition system: A second report. In D. LaBerge and S.J. Samuels (Eds.). Perception and Comprehension. Hillsdale, N.J.: L. Erlbaum Assoc., 1978.

[7] Anderson, J.R., Kline, P.J., & Beasley, C.M. A theory of the acquisition of cognitive skills. In G.H. Bower (Ed.). Learning and Motivation, Vol. 13. New York: Academic Press, 1979.

[8] Anderson, J.R., Kline, P.J., & Beasley, C.M. Complex Learning. In R. Snow, P.A. Frederico, & W. Montague (Eds.). Aptitude, Learning, an Instruction: Cognitive Processes Analyses. Hillsdale, N.J.: Lawrence Erlbaum Assoc., 1980.

[9] Anderson, J.R. & Kline, P.J. A learning system and its psychological implications. To appear in the Proceedings of the Sixth International Joint Conference on Artificial Intelligence, 1979.

[10] Reder, L.M. & Anderson, J.R. Use of thematic information to speed search of semantic nets. Proceedings of the Sixth International Joint Conference on Artificial Intelligence, 1979, 708-710.

[11] Neves, D.M. & Anderson, J.R. Becoming expert at a cognitive skill. To appear in J.R. Anderson (Ed.), Cognitive Skills and their Acquisition. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981.

[12] Anderson, J.R., Greeno, J.G., Kline, P.J., & Neves, D.M. Learning to Plan in Geometry. To appear in J.R. Anderson (Ed.), Cognitive Skills and their Acquisition. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981.

E. Funding Support:

A Model for Procedural Learning,
John R. Anderson, Principal Investigator,
Office of Naval Research (N00014-77-C-0242)
$175,000 September 1, 1978 - September 30, 1980

II. Interaction With the SUMEX-AIM Resource

A. & B. Collaborations, interactions, and sharing of programs via SUMEX.

We have received and answered many inquiries about the ACT system over the ARPANET. This involves sending documentations, papers, and copies of programs. The most extensive collaboration has been with Greeno and Lesgold who are also on SUMEX (see the report of the Simulation of Comprehension Processes project). There is an ongoing effort to assist them in their research. Feedback from their work is helping us with system design.

We find the SUMEX-AIM workshops (those that we could manage to attend) ideal vehicles for updating ourselves on the field and for getting to talk to colleagues about aspects of their work of importance to us.

Due to memory space problems encountered by ACT we expect that soon we will need to make use of the smaller version of INTERLISP developed at SUMEX for use in the CONGEN program.

C. Critique of resource management.

The SUMEX-AIM resource has been well suited for the needs of our project. We have made the most extensive use of the INTERLISP facilities and the facilities for communication on the ARPANET. We have found the SUMEX personnel extremely helpful both in terms of responding to our immediate emergencies and in providing advice helpful to the long-range progress of the project. Despite the fact that we are not located at Stanford, we have not encountered any serious difficulties in using the SUMEX system; in fact, there are real advantages in being in the Eastern time zone where we can take advantage of the low load on the system during the morning hours. We have been able to get a great deal of work done during these hours and try to save our computer-intensive work for this time.

Two location changes by the ACT project (from Michigan to Yale in the summer of 1976 and from Yale to Carnegie-Mellon in the summer of 1978) have demonstrated another advantage of working on SUMEX: In both cases we were back to work on SUMEX the day after our arrival.

III. Research Plans (8/80-7/86)

A. Project goals and plans:

Our long-range goals are: (1) Continued development of the ACT system; (2) Application of the system to modeling of various cognitive processes; (3) Dissemination of the ACT system to the national AI community.

Our more immediate goals (for the next year or two) involve application of the ACTF system, whose development we have finished, to three domains. First, we hope to complete the development of a simulation of geometry learning in the system. Second, we are starting to embark on an effort to model the acquisition of programming skills in LISP. This will serve as another test of the ideas that we have developed in geometry about learning and planning. The third application will be the modelling first language acquisition. This is a more radical departure from our work in problem-solving and so will provide a rather different test of the learning theory.

B. Justification for continued use of SUMEX:

Our goal for the ACT system is that it should serve as a ready-made "programming language" available to members of the cognitive science community for assembling psychologically-accurate simulations of a wide range of cognitive processes. Our intention and ability to provide such a resource justifies our use of the SUMEX facility. This facility is designed expressly for the purpose of developing and supporting such national AI resources and is, in this regard, clearly superior to the facilities we have available locally from the Carnegie-Mellon computer science department. Among the most important SUMEX advantages are the availability of INTERLISP on a machine accessible by either the ARPANET or TYMNET and the existence of a GUEST login. It appears that, at least for the time being, ACT has no hope of being a national resource unless it resides at SUMEX and, given the local unavailability of a network-accessible INTERLISP, it would even be very difficult to shift any significant portion of our development work from SUMEX to CMU.

C. Needs and plans for other computational resources

Carnegie-Mellon's plans to begin upgrading its PDP-10 hardware to emerging state-of-the-art machines (VAX, LISP machines, etc.) promises to provide a excellent resource eventually, and we hope to have access to that resource as it develops. However, given that a considerable amount of software development will be required, a sophisticated LISP system such as INTERLISP is not likely to be available on this hardware in the near future.

D. Comments and suggestions for future resource goals:

We are beginning to feel squeezed by various limitations of the SUMEX facility. The problem of peak load is quite serious. We have also been struggling with the address limitations of the current INTERLISP which is made more grievous by the amount of space INTERLISP requires. The computation time and address space limitations have meant that we have not been able to pursue certain projects that we would have otherwise. We applaud any efforts to increased computational power, to increase the address space of INTERLISP (e.g. VAXes), or to create significantly more space efficient versions of INTERLISP.